DOCUMENT RESUME

ED 073 777 LI 004 187

AUTHOR	Heilprin, Laurence B.
TITLE	Impact of the Cybernetic Law of Requisite Variety on a Theory of Information Science.
INSTITUTION	Maryland Univ., College Park. Computer Science Center.
REPORT NO	TR-236
PUB DATE	Mar 73
NOTE	52p.; (20 References)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Cybernetics; *Information Dissemination; *Information Science; Information Systems; *Information Theory; Models; *Theories
IDENTIFIERS	*Scientific and Technical Information

ABSTRACT

Search for an integrated, comprehensive theory of information science (IS) has so far been unsuccessful. Appearance of a theory has been retarded by one central constraint, the large number of disciplines concerned with human communication. Crossdisciplinary interdependence occurs in two ways: theoretical relation of IS phenomena to a given science, and practical relation of the science to IS dissemination of its contributions. We are concerned here with the first. The main difficulty in making IS models is the above mentioned interaction of many sciences of communication. Prominent among these are physics, biology, psychology, library science, computer science, several social sciences, applied logic and mathematics, and not least, cybernetics. That cybernetics should apply to IS is not a new idea. But more precision has emerged recently. Cybernetics is now seen to underlie nearly all IS phenomena through two central concepts: variety, and the law of requisite variety. The last provides a quantitative approach to system regulation and control. IS concerns regulatory activity in typically large, goal-seeking systems. That is, propagation of meaningful human messages between sender and recipient on an "IS path" is a goal-seeking process. (Author)

Technical Report TR-236 March, 1973

IMPACT OF THE CYBERNETIC LAW OF REQUISITE VARIETY ON A THEORY OF INFORMATION SCIENCE* **

Laurence B. Heilprin***

* paper given at Symposium: Perspectives in Cybernetics, arranged by American Society for Cybernetics at the 139th Annual Meeting, American Association for the Advancement of Science, Washington, D.C., 26-27 December, 1972.

**This article extends certain remarks on the occasion of the first session of a new special interest group on foundations of information science (SIG/FIS), American Society for Information Science, Annual Meeting, Washington, D.C., October, 1972.

***Professor of information science, School of Library and Information Services, and Computer Science Center, University of Maryland

ABSTRACT

Search for an integrated, comprehensive theory of information science (IS) has so far been unsuccessful. Appearance of a theory has been retarded by one central constraint, the large number of disciplines concerned with human communication. Crossdisciplinary interdependence occurs in two ways: theoretical relation of IS phenomena to a given science, and practical relation of the science to IS dissemination of its contributions.

We are concerned here with the first. The main difficulty in making IS models is the above mentioned interaction of many sciences of communication. Prominent among these are physics, biology, psychology, library science, computer science, several social sciences, applied logic and mathematics. and not least, cybernetics. That cybernetics should apply to IS is not a new idea. But more precision has emerged recently. Cybernetics is now seen to underlie nearly all IS phenomena through two central concepts: variety, and the law of requisite variety. The last (stated by Ashby) provides a quantitative approach to system regulation and control. IS concerns regulatory activity in typically large, goal-seeking systems. That is, propagation of meaningful human messages between sender and recipient on an "IS path" is a goal-seeking process.

The IS path is biophysically and topologically invariant in space and time. It consists in three segments, two within the communicants and one external to both. Within this environmental segment are located the physical artifacts of IS. All social structures depend on variation and matching of the central segment. This includes transmission of knowledge and education. The law of requisite variety affects each stage of propagation along an IS path. More specifically, it has shaped all steps in the bibliographic access to knowledge. Applications of the law of requisite variety at interfaces in IS systems are described in references to papers by the author and his students.

A quarter of a century ago cybernetics was founded by Wiener, Shannon, Ashby, Beer and many others. In the interval it has become a well developed science. Wiener coined the word "cybernetics" (from the Greek: "steersmanship") to designate the science of "control and communication in the animal and the machine".[1]

Fifteen years ago the phrase "information science" was coined to designate an anticipated science looming behind an immense agglomerate of scientific, technical and professional knowledge concerning propagation of human messages. The emphasis on communication overlaps that of cybernetics, and includes all aspects of the way messages are generated, transformed, stored, disseminated, received, and used. The field has far-reaching practical applications (one of which is communication in science). So far, however, the anticipated theoretical science has not appeared. Few or no principles have been found that integrate the vast array of information scientific phenomena. On the other hand, since these concern human thought and its communication it seems likely that "IS theory" will be related to scientific epistemology. This impression may be strengthened by what follows.

An aim of this paper is to present a framework for information science--an exhaustive domain containing (but not describing) all possible IS phenomena. A second is to demonstrate that cybernetics and information science are inseparably interrelated, and that at least one relation provides IS with an initial unifying theory. The relation has an epistemological implication that extends beyond IS to all science. In carrying this out, a problem arises: how to present a rather extensive but needed conceptual background to this distinguished but heterogeneous audience, within the brevity of a paper. The plan is to sketch several main concepts from cybernetics, describe the information scientific models, and demonstrate the applicability of concepts to models through selected examples. The cybernetic concepts are variety, systems (which are the "objects" that show variety), system transformations, system stability, and system regulation for stability or "survival" through the law of requisite variety. Much of this material is based upon Ashby's classic work.[2] The information scientific models are the author's "IS path" and "IS domain".[3-6] The IS path is a physical/psychobiological model of two communicating humans--similar to but not the same as Shannon's model of a communication system.[7] Some of the examples are from recent papers by the author's students.

This article is divided as follows:

I Variety and constraint

II Systems, transformations, stability

III Requisite Variety

IV IS models

V Application of I-III to IV.

I. Variety and Constraint

1. Variety is one of a number of concepts concerned with "quantity of information".[8,9] It is a property of a set, not of an individual. Variety is simply the number of discriminable differences which an observer can make in observing some system. We say that a system "shows variety". Since the discrimination is made by the observer, the variety in a set may be more (or less) for one observer (or discriminating system) than for another. Variety can be measured in many ways, the two most common being linear and logarithmic. That is, variety is measured either by the number N of values or states distinctly discriminated, or by the logarithm of this number, $\log_A N$, where A is some arbitrary base (usually 2):

$$V = N \text{ , or } V = \log_2 N \text{ ,}$$

$N = 1,2,\ldots,N_m$, where N_m is the maximum number of states or values observable under ideal conditions. The variety in the following letter:

Dear Dad: Please send money. Love. Your son,

is

12 (or $\log_2 12 = 3.58$ bits) for the set of <u>letters</u> in the message;

8 (or $\log_2 8 = 3$ bits) for the set of <u>words</u> in the message;

1 (or $\log_2 1 = 0$ bits) if the set is the <u>message as a whole</u>, or the <u>unit</u> set,

assuming we all count and discriminate identically.

The example shows that any observed object or system can give rise to different sets. Psychologically, the above observer "defines" three sets by paying attention respectively to three levels of substructure. First he observes the system the most closely, in its finest detail. In doing so the semantic aspect or meaning of the individual letter symbols is in the center of awareness, while the meanings of their groupings into words is not, or is less so; and even less the total purport of the structure of words in the message. In focussing on the words, he defocusses or gives up central attention to the letters, but gains some peripheral clarity or greater awareness of the entire message. Finally, in thinking of the message as a whole the meanings of the individual words are defocussed, merged and partly lost, while that of the letters is even more deeply suppressed. There is an analogy here with viewing a microscope slide at three decreasing stages of magnification, or with walking away from an object and pausing at three different distances, to look back at it. The object gives rise to three distinct sets.

Another common way in which an object can give rise to different visual sets occurs when we do not change the magnification but "scan" the object in a sequence of observations. In the background is retained awareness of the fact that the "object" seen in each view is part of a larger whole. However, in no two "frames" does the observer see the <u>same</u> object; actually he sees different substructures. Reading a page in a book is a common example.

2. An important case arises when the <u>same</u> system gives rise to different quantities of variety. This can happen if, for any reason, full observation is impaired or the variety is "constrained". Constraint on the variety of a set may arise externally in the system itself (some of a bank of switches work, others do not); or in the communication channel (the telescope diaphragm cuts off some of the light); or internal to the observer (he may be color blind). The concept of constraint is fundamental in all science, whether based mainly on observation or on reasoning. Information theory and cybernetics are founded on it. Just as variety is the number of discriminations the observer can make in some system, constraint on variety

is a relation between two systems. In this case the two systems are actually the same system observed under different conditions. The constraint is expressed by the difference between the full variety (observed under ideal or theoretical conditions) and the variety observed. The absolute constraint is simply the difference

$$C_{abs} = V_m - V$$

in which V is the variety observed and V_m the maximum variety observable if the set is not under constraint. Relative constraint would be

$$C_{rel} = (V_m - V) / V_m = 1 - V/V_m \quad .$$

Using the same definitions of constraint the linear and logarithmic measures of variety for the same range ($N = 1,2,\ldots,N_m$) do not give identical values. For example if a bank of three lights each of which can be on or off (and therefore ideally capable of $2^3 = 8$ discriminable states) were never used to display more than four states, then the absolute and relative constraints for the linear measure would be 8-4 = 4 and 1-4/8 = 1/2; while for the logarithmic measure the constraints would be $\log_2 8 - \log_2 4 = \log_2 8/4 = 1$ bit, and $1 - \log_2 4/\log_2 8 = 1 - 2/3 = 1/3$. Evidently if $V=V_m$ then with perfect freedom to vary, or "independence", both absolute and relative constraints have value 0. However, when the variety is reduced to 1 in the linear measure and 0 in the logarithmic (no freedom to vary), the values of the absolute constraints in the two systems would be N_m-1 and $\log_2 N_m - \log_2 1 = \log_m N_m$, and the relative constraints would be $1 - 1/N_m$ (nearly but not quite 1), and $1-\log_2 1/\log_2 N_m = 1$. The two measures of variety could be made to give identical relative constraints 0 and 1 at these intersections of the linear and log constraint curves if we were willing to define linear variety V = N-1. They would still not agree elsewhere.

3. Now, if the reader has gained the impression that variety and constraint on variety are rather simple concepts, let him not also assume them insignificant. On the contrary, simplicity endows them with a breadth of applicability unmatched in science. Perhaps we can put it this way: in the mental world variety is comparable to atoms. The physical universe is built of atoms, the mental universe is built of variety. Consider the following: (1) that variety is more general than its closely related concept, entropy; (2) that constraints on variety form the bases of all natural laws and formal axiomatic systems;[10] (3) that processes such as recollection, learning, teaching, predicting, deciding -- all cognitive and evaluative functions -- depend on structures of variety and its constraints;[11, 12] (4) that variety and constraint retain their usefulness for cognitive and evaluative processes at all levels, i.e., are "invariant" under, or independent of, the transformations of abstraction;[12] (5) that the law of requisite variety forms the basis of regulation and control of systems.[13, 14]

While expressions of constraint tend to be quite different in each discipline, of particular interest is the way in which entropy expresses constraint on a set in information theory:[15, 16]

$$H = -\sum_{j=1}^{n} P_j \log_2 P_j$$

in which P_j is the probability of occurrence (in a long sequence) of the variety in the jth state, given n states. Entropy measures an average, the average of the 'information' $-\log_2 P_j$ in the jth state, over all states $(j = 1,2,\ldots,n)$. When there is no prior knowledge of any particular outcome all outcomes are equally likely and $P_j = 1/n$, for all states j. As is well known the expression for entropy becomes

$$H = -\sum_{j=1}^{n} 1/n \log_2(1/n) = n/n \log_2 n = \log_2 n$$

and this is also the maximum value taken by the entropy. The condition for maximum entropy is identical with the condition for least knowledge about the set of n states or values, i.e., for independence of probabilities of occurence of the outcomes, for "absence of constraint", not on the <u>occurence of variety but on the relative frequency of occurence</u>. It is also the condition for equivalence between Hartley's original measure of "information capacity" and Shannon's entropy[16, 17]; and finally, it is equivalent to the logarithmic measure of variety in the set.[18] The logarithmic measure of variety can be regarded <u>either</u> as a simplification to which the measure of entropy of a set showing variety is reduced when additional knowledge about the states (frequency of occurrence) is least (i.e., entirely lacking) <u>or</u> as a measure of <u>something more general</u> than entropy. Entropy requires for its specification additional knowledge, knowledge of the relative probabilities of the different states of variety. Hartley's information capacity was conceived as the measure of the number of different units (whose individual capacity is shown by the log base) needed for physically storing a "representation" of a message of n distinct symbols, without regard to possible constraints of which one could take advantage. That is, storage capacity takes no advantage of the relative frequencies of occurrence of the symbols, in order to recode the message to a lower capacity. The minimum recoded capacity was later defined as the "information content" to distinguish it from "information capacity". Variety, information capacity, maximum information content, and maximum entropy are numerically equal.

We now see that entropy need not enter a discussion of variety unless we require not one but two kinds of information. Variety is a more primitive concept based only on discriminability and not also on relative frequency of occurrence of what is discriminated. We need entropy for, e.g., efficiency of coding messages into a given channel or space. We do not need it for an inquiry into the applicability of the law of requisite variety to information science. In fact, the general applicability of this law to all forms of regulation might be reduced if we had to know too much in order to use it.

4. The discussion could be prolonged, for variety is related to fundamental thought processes. We confine ourselves to two comments. First, there is an apparent inconsistancy between "operational" definitions of the measures of variety and constraint on variety. The former is defined as the number of discriminations in relation to the observer--those he or she discriminates. If for any reason the number is less than the full variety the observer cannot necessarily be expected to know this. How define a constraint (which requires knowing the full set, N_m) that one cannot observe? The inconsistency is removed if one or more other observers (who represent "objective" scientists measuring the constraint) can observe both the unconstrained and the constrained sets; or if the same observer <u>can</u> in some way determine how much less he observes than he might observe. The formulas are therefore valid for one observer who can subjectively compare the two sets, or for two or more observers who objectively determine the full variety and the variety accessible to the observer. The writer has suggested elsewhere that all observation is, strictly speaking, subjective; and that what we call objective is simply a subset of the subjective more completely checked by intercommunication.[19]

The second comment is to call attention to the extreme versatility of the concept variety. Logically, we could better understand this after discussing the law of requisite variety, for that law expresses relative versatility in coping with difficulties. However, two simple examples will show what is meant. The variety in color discrimination is a sensory experience at the most concrete level, that of the direct sense impression or DSI. The variety in a set of logical outcomes is abstract. Consider, for example, the conditional proposition on which are based the hypotheses of science: if A then B. The conditional is a single statement compounded from two elementary statements A and B, each of which can have two attributes (or states), true or false. The truth table is

A	B	$A \rightarrow B$
T	T	T
T	F	F
F	T	T
F	F	T

In other words the conditional defines an abstract set, all the possible outcomes which can occur, given the four hypothetical combinations (TT, TF, FT, FF). As long as there are no restrictions on the possibility of occurrence of any combination of states of A and B, the set displays full variety: 4 or $\log_2 4 = 2$ bits. If, however, by any independent means it is possible to add the observation that the state (TF) never occurs, then the variety in the truth table is reduced from four outcomes to three. The proposition is now universally true (all T), and the conditional has been converted into an implication: A implies B. This is a set relation, a new property acquired by the truth table set by reason of the constraint on its variety.[20]

The example demonstrates that in logic and mathematics, far from the concrete end of the spectrum of abstraction, the concepts of discriminable states, variety, and the formation of relations through constraints placed upon the full set of states are applicable. Whether we are dealing with a discriminable set of direct sense impressions or with an abstract set, we can measure the variety of the set, the constraints on the set, and the resulting relations on which our rational thought depends. More vividly we might say that discrimination and counting discriminations are invariant under the mental transformations of abstraction.

II. Systems, Transformations, Stability

1. The concept of a system is so widespread that it is used in all science. What are the common characteristics of systems? The author has suggested that a system is an abstraction from a real object:

> The reason for including our subject within broader classes is simple. Human thinking is performed in terms of classes, and a system of any kind is simply a class located within a larger class--its environment. The processes of the system may be divided into what have been called its "internal transformations", and those processes which cross boundaries and relate it to its environment.[21]

Since a class is mental, a system is a mental construct, changeable at the will of the conceiver. We usually draw a mental boundary around the system, such that internal and external parts, and relations between parts are separable. We feel no compunction in shifting attention to a part or subsystem, and calling <u>that</u> the system. This is precisely what we did in the case of the example of the different sets to which an object or system can give rise. Thus systems can form ordered hierarchies of sets, and mathematically can be treated abstractly, e.g., by lattices.

The above is perfectly general, and applies to all systems--they retain the trace of their origin by always having an environment-- the distinguishing feature of a real object. Real objects also change, and all except the most trivial systems are subject to transformations. The system is said to be in some 'state', fixed by a set of values of the variables which together define the system; or often, simply by a single symbol which defines the state. For simplicity, using only the latter, consider the system whose states are given by set {A, B, C, D} . Its transformation T is representable as follows:[22]

$$T \downarrow \begin{matrix} A & B & C & D \\ B & C & D & A \end{matrix}$$

We call the initial states (top row) the 'operands', the final states the 'transforms'. The transformation T is also shown graphically in Fig. 1a. The transformation shows that if the initial state of S is A, at the next "step" S will be at state B. If the transformation is repeated twice, we can say that the transformation T^2 takes the system S from A to C, T^3 from A to D, and so on. If the system is a model of a real object (has a real referent for the mental construct) then it will take a finite time t_{AB} for the system to be transformed from A to state B, another interval t_{BC} for the second transformation, and so forth. For simplicity let us assume that all the steps take an equal time t, and that the transformation is repeated as soon as each 'step' is completed. Then system S will go through all its states in succession, and cycle back over them, indefinitely. Given the initial state A and n steps (or applying T^n) we can predict exactly in which state system S will be at the end of time nt (e.g., if $n = 5$, the system will be at B). Suppose, now, the time interval nt represents real time. Then we have a way of schematically representing how any real object passes in succession through its various states. More generally, we can think of the system S as being the real world about us, transformed stepwise (each moment of perceptible duration) into the real world of the next moment--some of it remaining unchanged, other parts of it changed into new states. The steps can be short enough so the change seems continuous. Transformations are the formal way of representing either changes in real time of real objects, or purely hypothetical changes.

Now suppose that system S has a more complex structure represented by the following _set_ of transformations:

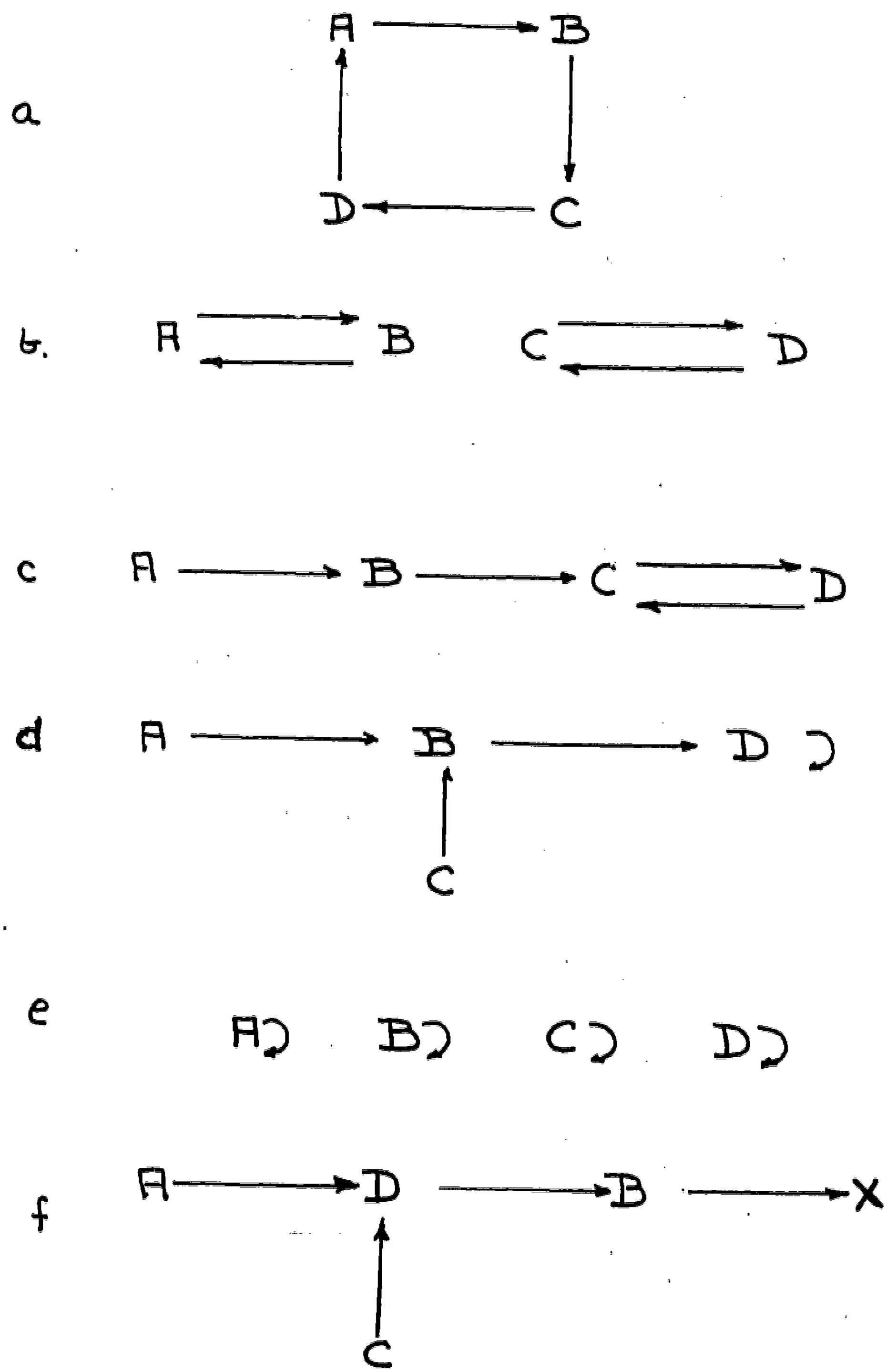


FIG 1 SYSTEMS AS MACHINES

T	↓	A	B	C	D
	a	B	C	D	A
	b	B	A	D	C
	c	B	C	D	C
	d	B	D	B	D
	e	A	B	C	D
	f	D	X	D	B

This represents a "machine with input".[23] Each row represents a possible transformation on the four states A, B, C, D. In row a, the machine transforms as before, in only one manner--shown in Fig. 1a. This is its internal mode of operation as long as nothing changes the system. Now the only thing that can change system S is not <u>inside</u> it, for S is determined by its transformation ("internal constraints") T. Therefore, the only possible agent which can change S from its mode of function in row a is some <u>external</u> agency--in the environment. If something in S's environment shifts the "input" parameter from a to b, then S now transforms according to the new mode of row b, shown in Fig. 1b. The input from the environment is said to control the states to which S transforms, these states being S's output. That is, we think of the system S 'coupled' to the environment so that the 'supersystem' of S and its environment forms one machine. The environment 'drives' S in the sense that it selects the parameters (a,b,...,f) i.e., not the initial state of S but the transformation. Again, the output states of S can be the input to a second system, U coupled to S, such that the transform states of S are the parameters which select the internal transformations of system U . In this way we can model intricate relations between environment and systems, or between subsystems of any system, coupling them through their inputs and outputs. We can go on to design model systems of components subsystems which perform according to actual or theoretical requirements. We can test the operations of these models by the use of the coupling transformations and the internal transformations.[24] Thus cybernetics is a "science of the artificial" in the sense that its chief concern is system analysis and design (for control) of systems, or artifacts.[25]

Returning now to our extended set of transformations T of machine S with inputs a-f , we know of course that with four states and no constraints there could be 4^4 or 256 different modes of internal operation. The fact that there are only six is therefore a large constraint on the variety shown by S . The constraint defines the system S . The first five rows show a few characteristic types of internal transformation. In a , there is one "basin", with a complete cycle; in b two basins, and two cycles; in c one basin with a terminal cycle; in d one basin and a terminal state at which the system comes to "rest" in the sense that it transforms identically into the same state D after, at most, two steps. In e the four states transform identically into themselves such that if the system starts in any state it remains there. All rows represent single valued transformations, in the sense that no state transforms to ("maps into") more than one state. "Single valuedness" is the absolute minimum we can require of a "determinate" machine, i.e., one whose states can be predicted. (Note: stochastic machines, not discussed here, retain some predictability). The machine, however, can be determinate without resembling our usual idea of a machine, which must continue to operate. This added feature requires a closed transformation, one in which no new states appear in the transforms that were not present among the operands. The variety of internal states must not increase. Rows a through e are closed, and represent our idea machines (single valued, closed transformations). However, these five machines (or five modes of operating one machine) are quite distinct from eachother. The machine represented by row a cycles, and retains all its initial variety. In addition to being single valued and closed, its transformation is 1-1 (as many different transform states as operand states). Row e also preserves its variety. But here, if the system starts in any operand state it remains there. An observer would not discover the other states either before or after transformation, and would not know it had inter-state variety. In row b only two states of the four would be observable; in row c states A and B might be observed but would never recur after the second step--after this step they would be "inaccessible". The same would be true for row d after the second step. We conclude that, with a finite number of states, a system that is repeatedly transformed can at most retain all its initial variety. It does so only if the transformation is 1-1 at each step. Much more often the transformation is many-1 for at least a few steps, and the variety shown by the states decreases. If the variety in a machine is used to code messages, after repeated transformations the information in the messages is at most preserved, and usually decreases.

2. System stability is closely related to system transformations. Finite state systems such as that in Figure 1 either come to equilibrium at a single state (rows d and e) or to a terminal cycle of states which we might call 'cyclic equilibrium' (rows a,b,c). A stable system is simply a system that, after displacement from an equilibrium state

or cycle to some other state, eventually returns (under the original transformation) to the original equilibrium state or cycle.

In order to displace a system in equilibrium, since the system itself cannot do this the displacement must come from without. The displacement therefore occurs through a shift to a new row of transform states, and the starting state transforms accordingly.

Consider the following single row transformation of a system with four states:

$$T \quad \downarrow \quad \begin{array}{cccc} A & B & C & D \\ A & A & D & D \end{array}$$

As shown in Fig. 2a, this system has two basins and two states of equilibrium. It can represent a machine, since T is single valued and closed--even though the machine stays quietly either at state A or state D after one step.

Now let the system be displaced in one step. Three different displacements (D_1, D_2, D_3) are shown below, with T above each:

↓	A	B	C	D	↓	A	B	C	D	↓	A	B	C	D
T	A	A	D	D	T	A	A	D	D	T	A	A	D	D
D_1	C	B	D	A	D_2	D	C	A	B	D_3	D	X	D	B

Suppose the starting state is B. The one-step transformation D_1 is $D_1(B) = B$, and the next step (under T again, after displacement D_1 ceases) is $TD_1(B) = A$. Therefore the system returns to the same state of equilibrium (A), and is stable under D_1. If the starting state had been D, however, the system would regain equilibrium, but in the other basin, at A. That is, $TD_1(D) = A$. Therefore if the starting state is D the system after displacement D_1 is still a machine, still in equilibrium, but not stable.

a. T $\quad (A \longrightarrow B \qquad C \longrightarrow D)$

D_1

$$\begin{tikzcd} A \arrow[r] & C \arrow[dl] \\ D \arrow[u] & \end{tikzcd} \qquad B$$

b. D_2

$$\begin{tikzcd} A \arrow[r] & D \arrow[d] \\ C \arrow[u] & B \arrow[l] \end{tikzcd}$$

D_3

$$\begin{tikzcd} A \arrow[dr] & & & \\ & D \arrow[r] & B \arrow[r] & X \\ C \arrow[ur] & & & \end{tikzcd}$$

c.

$TD_1(B)$	$B \Rightarrow B \rightarrow A$	$=, S$
$TD_1(D)$	$D \Rightarrow A$	$=, \bar{S}$
$TD_2(B)$	$B \Rightarrow C \rightarrow D$	$=, \bar{S}$
$TD_2(D)$	$D \Rightarrow B \rightarrow A$	$=, \bar{S}$
$D_3(B)$	$B \Rightarrow X$	
$TD_3(B)$	— — — —	$!, \bar{S}$
$TD_3(D)$	$D \Rightarrow B \rightarrow A$	$=, \bar{S}$

$\Rightarrow$ CHANGE OF STATE, DISPLACEMENTS D_j $\quad j = 1, 2, 3$

$\longrightarrow$ " " " , TRANSFORMATION T

$=$ IN EQUILIBRIUM $\quad S$ STABLE

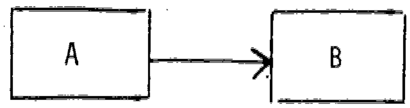
$!$ DISCONTINUOUS $\quad \bar{S}$ UNSTABLE

FIG. 2 SYSTEMS AND STABILITY

$D_2(B) = C$, and $TD_2(B) = D$; while $D_2(D) = B$, and $TD_2(D) = A$ so that we have the same inversion as above. Here, however, no matter what starting state is used, the system is unstable.

D_3 has the same row as f , in Fig. 1f. And $D_3(B) = X$, a new state. The system cannot transform again according to T , and there is no recovery. Interpreting this as the operation of a real machine, the machine has passed into a state not only of "no return" to its initial state of equilibrium but of impossibility to continue. Its existence as a "machine" ceases. Thus we associate the closed transformation with "continuity" or survival of a system; and a single valued, open transformation is a machine only in the sense that it may start off, perform one or more steps, and then permanently stop. If a system stops at a state or cycle of equilibrium it can be "started" again by external means. But if it stops in a new state there is no way, even for an external agent, to get it back into a state from which it can continue to perform. It has undergone an irreversible, final transformation. Picturesquely, we can say that the system is destroyed. If the system represents a living organism or organization, it is dead.

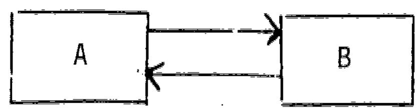
3. In the directed graphs ("digraphs") of Figures 1 and 2 a vertex (letter) represents a state, and an arc represents the transformation. The arc is directed from the operand to the transform state. In another form of digraph called a "diagram of immediate effects" vertices (in boxes) represent parts of coupled systems if connected by arcs, independent or "reducible" parts if not. The arc direction shows output (arc leaving) and input (arc entering). Part A whose output is the input to part B



A dominates B

A and B are independent

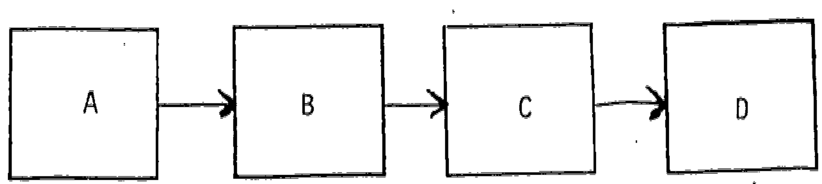
selects the rows of B's transformations and regulates, controls or "dominates" B. If there is also an arc from B to A



A and B are symbiotic

the two subsystems operate by "feedback". Each "lives" symbiotically in the environment provided by the other. For a fully-connected digraph of more than two vertices the concept of simple feedback becomes inappropriate. It is intuitively apparent that parts of systems harmoniously (non disruptively) coupled by reciprocal inputs and outputs must synchronize their internal transformations so that all transformation times are integral multiples of each other (a natural time quantization). Further, if one part is in cyclic equilibrium, so must be the other, otherwise it would "veto" ("disrupt") the internal transformation of the part or parts it dominates (e.g., a strike by bus drivers, or longshoremen).

If several vertices (A,B,C,D) are coupled into a chain so as to form a "communication channel"



it may take two or more steps (or time units) for the actual dominance of A over C and D to become observable. After three steps the above becomes a "diagram of ultimate effects".[26]

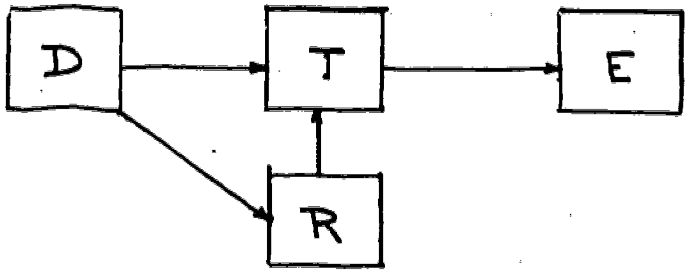
Possibly the main use for this kind of diagram is in connection with the law of requisite variety.

III. Requisite Variety

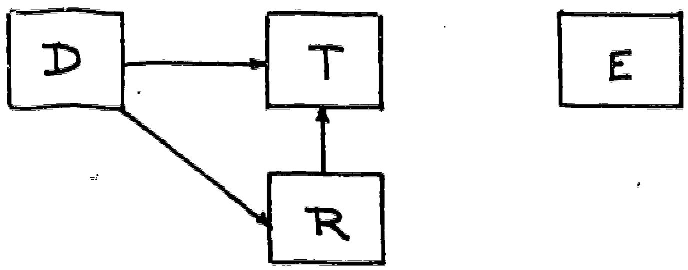
The cybernetic law of requisite variety combines several ideas already described. It is readily understood by a diagram of immediate effects with four main parts (E,D,T,R) coupled as shown (Figure 3a).

> Let E represent a set of "essential variables" which must be kept within certain limits in order to achieve a goal. Let D be a set of disturbances or threats to E which, by acting on E through the environment T, can drive E's variables out of the region of stability for attaining the goal. Finally, let R be a regulator interposed to keep D from disturbing E by driving its essential variables out of bounds. Then the law states that R can successfully "regulate" D only if it has "requisite variety". That is, R must have a variety of alternatives sufficient to counter the variety of disruptive alternatives open to D.[4]

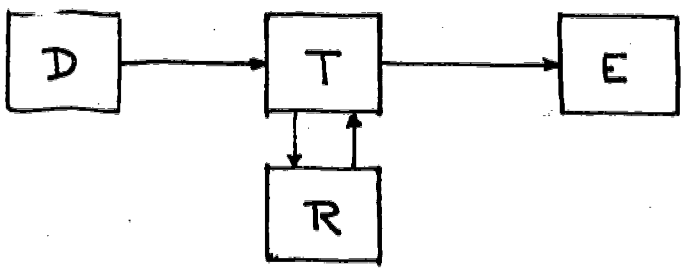
The discussion could be given in terms of states, or of variables. The latter may be preferable for visualizing limits. If an organism can survive within a temperature range (say) -20°F to +120°F and within a pressure range 3 lbs/in^2 to 45 lbs/in^2, such ranges represent the kinds of limits within which the organism (set of variables E) must be kept. If E is a machine, the machine has a like set of working range limits. These also can be viewed as ranges of stability. For continuation of E as a system is a goal fundamental to whatever other goals E may have. Thus we can partition the total set of values of E into E_+ (the subset or region of values for survival) and E_-(all other regions). The problem of regulating D is, simply, to keep D from driving E into E_-; or to keep E within subset E_+. If D is a "natural disturbance" it has no "objective" -- is not out to defeat E as would an opponent in a game. So some of D's acts may not harm or totally disable E. Nevertheless, in stating and presenting this central law of cybernetics (for naming and discovery of which he is credited) Ashby represents the situation as a contest between R and D, in a two person "matrix game". The elements of the matrix T are the outcomes



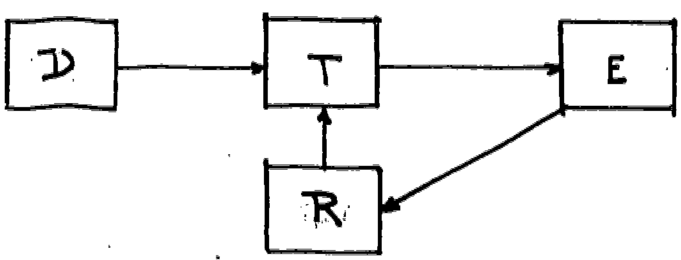
a 1-STEP COMMUNICATION CHANNEL REGULATION: D→R



b. PERFECT (1-STEP) REGULATION: D→R



c 2-STEP REGULATION: D→T→R



d 3-STEP ("ERROR CONTROLLED") REGULATION: D→T→E→R

FIG. 3 LAW OF REQUISITE VARIETY

of the joint interaction of D and R on E, acting through the environment represented by Table T of matrix outcomes. Some outcomes in T will leave E within E_+, free to continue, others within E_-.

Consistent with our picture of equilibrium (a displacement followed by a recovery to an E_+ state) the game consists in two moves:

> D plays first, by selecting a row of the matrix. R counters, by selecting a column. If the outcome (value of the matrix element jointly selected by D and R) is within E_+, then R wins, "regulating" D. Otherwise R loses. Only variety in R can "drive down" or "destroy" the variety in D.[21]

In Figure 3a regulator R receives notice of D's move directly, in time to make the best possible countermove in selecting a column of T. Replacing E_+ by "+" and E_- by "-", a game might look like this;

		R: 1	2	3	4
D	a	+	–	–	–
	b	–	+	–	–
	c	–	–	+	–
	d	–	–	–	+

Table T

With the particular table T shown, using only the first three columns (R's moves: 1, 2 or 3) and four rows (D's moves: a, b, c, or d), R can win if D plays rows a, b or c; and R loses if D plays row d. To make R into a "perfect regulator" such that E never experiences D's moves, we should have to increase R's variety. We could accomplish this by constructing a more versatile regulator R with four instead of three optional moves, such that there is a "+" on row d (dotted vertical lines).

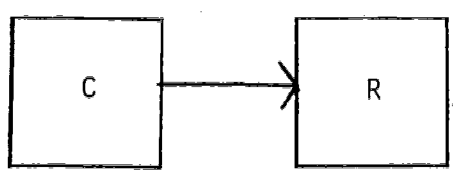
Obviously the situation changes with the structure of table T. In certain cases R can never win. In most cases R can reduce D's variety so that, while there may remain some possibility that D wins, the chance of immediate termination is reduced. The mathematical expression of the law of requisite variety takes the form of an inequality:

$$V'(D) \geq \frac{V(D)}{V(R)} \text{ (linear)}, \quad V'(D) \geq V(D) - V(R) \text{ (logarithmic)}.$$

This states that V'(D), the reduced or resultant variety which D can exercise, exceeds or at least equals D's original variety V(D) divided by R's variety V(R). In Ashby's words: "The variety in the outcomes, if minimal, can be decreased further only by a corresponding increase in that of R".[27] Obviously, if V'(D) = 0, (logarithmic measure) D has only one mode of affecting E, and R needs only one effective countermove. Even if V'(D) is a small number the regulation may still be perfect or nearly perfect. The essential is that, by reducing D's variety E's chances of survival are increased. This is the case with organisms and organizations in ordinary living. There is no perfect regulation (as in Figure 3b), no guarantee of continuity. All "live dangerously" -- i.e., contingent on the capacity of their regulators to reduce D's variety.

There is a systematic decrease in the amount of regulation R can provide, even with a favorable table T. In Figure 3a, as mentioned, the channel of communication by means of which R learns of D's move is 1-step, and we interpreted this as rapid enough so that D and R simultaneously act on T, and (if T is favorable) that it is possible for R to completely shield E from "knowledge" of D's move. In Figure 3b this was shown by "no message" getting through to E from D. However in Figures 3c and 3d the situation becomes progressively less favorable. In Figure 3c the path of communication needs two steps, and R can regulate D only <u>after</u> D has acted on the environment, but <u>before</u> the environment acts on E. Here the chance of R's selecting a "correct" move is impaired by the length of time delay -- and it becomes evident that "R's capacity as a regulator cannot exceed R's capacity as a channel of communication".[28] In Figure 3d R learns of D's move only <u>after</u> it has affected E. In this situation, that of the "error controlled" regulator, the best that R can do is let E suffer some losses (which are still in E_+ and not fatal) but which do indicate the threat; and then counter with a move that restores E to a better position in the "survival region". In other words, with the minimum 3-step channel R can <u>never</u> be a perfect regulator, but always an imperfect one. This type of partial protection predominates in nature -- continuity is contingent on developing better regulators. Businesses fail, empires fall, healths decline, species die out if disturbances cannot be regulated after-the-fact.

It is possible to add another part to the diagrams of Figure 3. This is a "control" with input to R



In this case R actually has two inputs, one from D, T or E and one from C. The effect of the inputs from D, T or E is to enable R to select a column in T. The effect of the input from C is to change R's input parameter -- its row. Thus C dominates R. This may work well, since it is an easy way to increase R's variety by giving it, not simply one row of column choices, but several rows. However, in the light of the fact that R's effectiveness depends on its capacity as a communication channel (an "error-correcting channel" is more precise[29]), if C does not also share the input to R (is not as informed about D as is R) the added mechanism (R and C acting as a unit with T to protect E) may not be more effective than the simpler mechanism (R and T acting together to protect E). In <u>regulation</u> R has a single row of options, while in <u>control</u> R has more than one row. This is analogous to the "machine with input". Control C dominates R's transformations, i.e., R's selection of a column, given a row selected by D.

Ashby makes the point (p. 214) that "perfect regulation of the outcome by R makes possible a complete control over the outcome by C ". What, then, is a perfect regulator? For R to be a perfect regulator two independent conditions must be satisfied. The following is one such pair:

1. R has "requisite variety" (a sufficient number of columns in Table T) so R can always select a favorable outcome (member of $\{E_+\}$), if it exists, in any row selected by D .
2. No row of T contains only unfavorable outcome elements (all members of $\{E - E_+ = E_-\}$).

Condition 1 can be said to depend on the quantitative aspect of variety. In this case it is the relative number of rows and columns in Table T. Condition 2 can be viewed as depending on the qualitative aspect of variety - in this case the intrinsic structure of Table T. Each row of T must contain at least one member of $\{E_+\}$. This shows us that the two conditions for perfect regulation cannot always be satisfied solely by design of regulator R . If any row of T exists such that R cannot select a favorable "target" in spite of its capacity to select any column, then R is an imperfect regulator.

The row condition, "invariant unfavorable outcome" (IUO), does exist for all real systems E . It may not be evident. A physical embodiment might be a row of elements corresponding to any of the following disturbing events (war, epidemic, nuclear explosion, extraordinary solar radiation, impact of a large meteor, planetary orbit dislocation with high index tidal waves and earthquakes, etc.). Any catastrophic "act of God" can produce an IUO. For these no foreknowledge or preselection by R or C can avail. Thus satisfaction of the law of requisite variety does not completely insure effective regulation. We must also satisfy the condition of at least one favorable outcome for each component of disturbance. This qualification in no way reduces the central importance for regulation of the law of requisite variety. It does show that in human and all biological affairs survival is contingent. IUO is always potentially present, however unnoticed. These observations can be summed up by a logical statement: satisfying the law of requisite variety is necessary but not sufficient for perfect regulation. Alternatively, system survival implies requisite variety in the system regulator, but the converse is not true.

IV. IS Models

These models have been discussed at various times.[3,4,5] The first, the "IS path" (Fig. 4), is a diagram representing schematically a one-way path of message propagation from an individual A to an individual B. Quoting References 3 and 4:

> The over-all path consists of three segments, each composed of many stages. Segment *ab* is in individual *A*, the author or sender of the message. Segment *bc* is in the environment or medium [or media] surrounding *A* and *B*. Segment *cd* is in individual *B*, the recipient of the message. The entire path is physical in the sense that all of its stages are subject to the laws of physics. We will call *bc* the external path segment (also "external segment," "external path," or "physical link"), *ab* the organic efferent path segment, *cd* the organic afferent path segment. Individuals *A* and *B* possess many peripheral sense organs and motor organs, but these are represented in each by a single effector or motor organ *M*, and by a single sensor or sensory organ *S*. The organ *M* is used to modulate suitable physical systems used as message carriers through intervening medium *bc*. The organ *S* is used for reception and transduction of the modulation within the same sense channel. However, it is not necessary that both *M* and *S* correspond to the same sense channel, provided suitable translaters and transducers exist. Each individual, *A* or *B*, is provided with a nervous system that conducts afferent modulation from peripheral sense organ *S* to central region *C*, which we call simply the "mind". It also conducts efferent modulation from *C* or a region near *C* to peripheral motor organ *M*. The exact locations of *M*, *S*, and *C* within the body boundaries are not material to our picture. But the fact that part of the path of propagation lies in each body is essential.
>
> It will be noticed that what the communication engineer usually thinks of as the path of propagation is external path *bc*. The stages of this part of the path extend from the boundary of the message sender through possibly many media to the boundary of the message recipient. These stages may have widely differing conditions of propagation. They may include natural media such as air, water, and solids; or man-made media such as transmitters, receivers, and information-storage-and -retrieval systems. *bc* may include organic, possibly living structures. But in general, the media outside the boundaries of *A* and *B* are purely physical and "non-semantic," in the sense that they do not contain the special message-initiating and message-receiving equipment located in the communicating organisms. Thus, relative to *A* and *B*, *bc* is simply a set of stages not including the bodies of *A* and *B*, through which the modulation passes without the special processing that occurs in nervous systems (later referred to as "association"). In *bc*, modulation remains invariant if it is propagated in the absence of noise. We assume the range of propagation to be such that the

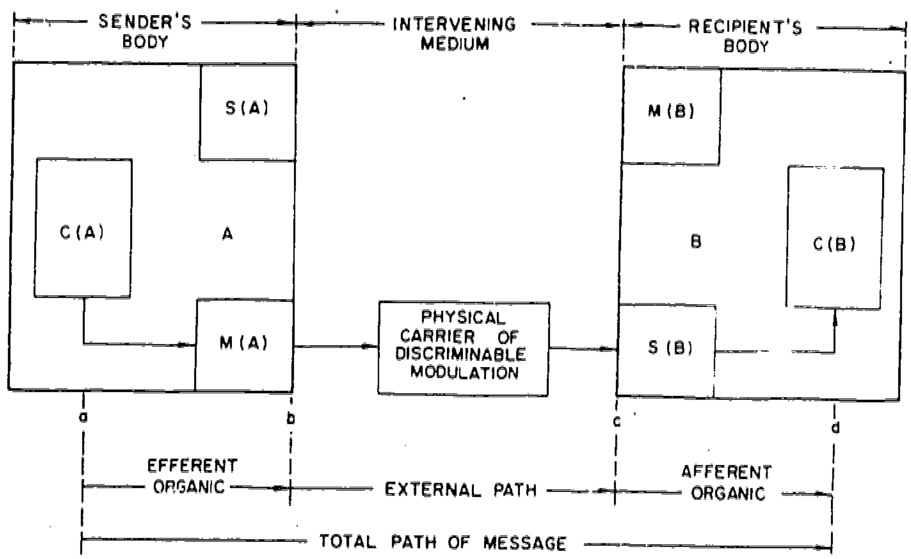


FIG. 4. PATH OF PROPAGATION OF A HUMAN MESSAGE (IS PATH)

power level remains high enough, and distortion low enough, for complete discrimination by *B* of the modulation encoded by *A*. The message may be amplified, regenerated, and transduced many times. But it remains simply modulation, physically transducible into itself in the form in which it left *A*'s modulating organ. With noise, the modulation deteriorates according to the second law of thermodynamics, and its "bit content" of information deteriorates according to the mathematical theory of communication. In many ways the purely physical modulated carrier in path segment *bc* is the simplest form assumed by a message as it passes from *A* to *B*.

Before generalizing this model, let us sketch in one more feature:

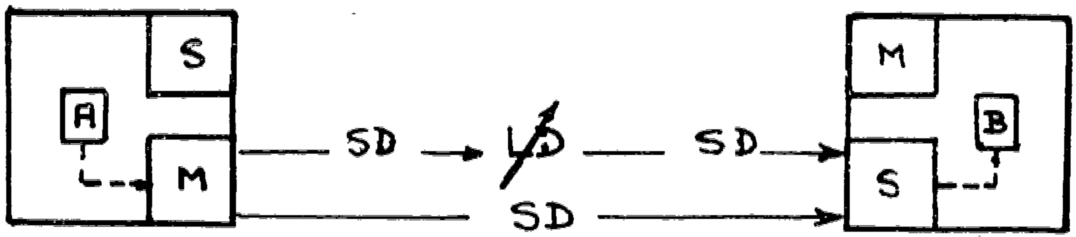
All messages may be divided into two classes: those of short duration (SD) and those of long duration (LD). Of these, SD are the more basic in the sense that we could communicate with only SD messages but not with only LD messages. A message in its simplest form consists of two components. The first is some physical system, which we will call a "carrier," that is not in itself a message (examples: radio waves without the voice or music "intelligence"; a blank sheet of paper). The second consists in discriminable marks on carrier such as images or sounds (SD) or printed letters or drawings (LD) which we will call "modulation". The carrier can exist without modulation but not the reverse. In SD messages, the modulation varies in time. The marks on sound waves in direct speech change constantly at the ear - in fact they must die away (be attenuated) rapidly, if we are to be able to discriminate the next words or musical notes. If they persisted for even a short time more than they do, the sounds of successive speech or music symbols would become indistinct and blurred... It is of the essence that SD messages be attenuated at least as fast as they pass into the sensor of a human recipient or of a machine-receiver. Somewhat more formally: the attenuation rate of the channel which conveys information to the sensor must equal or exceed the information rate of the sensor. By information rate is meant the time rate of change of fully discriminable "least units" of information such as word-symbols, or of their components, such as "bits". Unless we refer specifically to bits/second or other rate units, by "sensing rate" we shall mean "words/second".

The reason why SD messages are more basic than LD is simply that when the message passes into the sensor of man or machine it must do so in SD from. Human sensory (afferent) and motor (efferent) messages travel by time-varying modulation to and from their destination or source - usually the brain. The same is true of machines which pass information through a sensor into some "decision" mechanism.

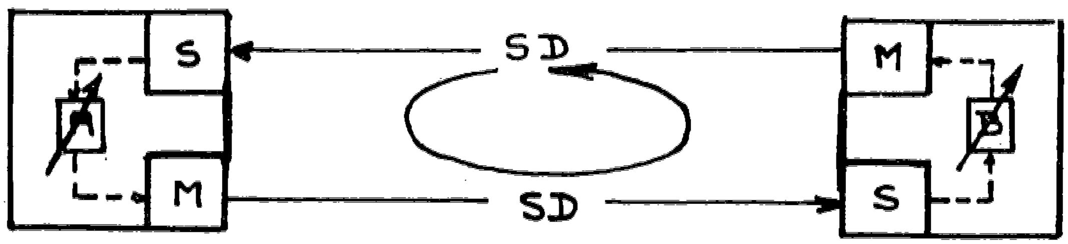
In contrast with that of SD messages, the modulation of LD messages persists for comparatively enormous time intervals. In order to achieve this extension into the time dimension, the carrier is restricted in most cases to a solid, and the modulation, instead of temporal, is spatial. Printed letters on a page store their contained message for long periods. They do so by extending spatially on the page. Naturally, since both SD and LD modulation exist in space and time, both are four-dimensional "marks." But the far shorter time duration of SD modulation makes us refer to it as "purely temporal," which it is not.... .

Because of this constraint on communication - that a message must enter the sensor in the SD mode - all "stored" or LD messages must be convertible to SD form. This is indeed the case. It is also true that many SD messages are convertible to LD mode, but this was not always so - and the conversion is man's peculiar discovery. He found that information in LD mode can be stored, i.e., propagated into the time dimension, even beyond the life of the message sender.[30] The discovery enabled the cumulation of knowledge - the possibility for man's finite brain to tap a much larger memory than his own.[31] Using this technique of storage as a tool, he erected science and civilization. But it is a basic constraint on use of the tool that stored messages be retransformed into the SD mode.[37]

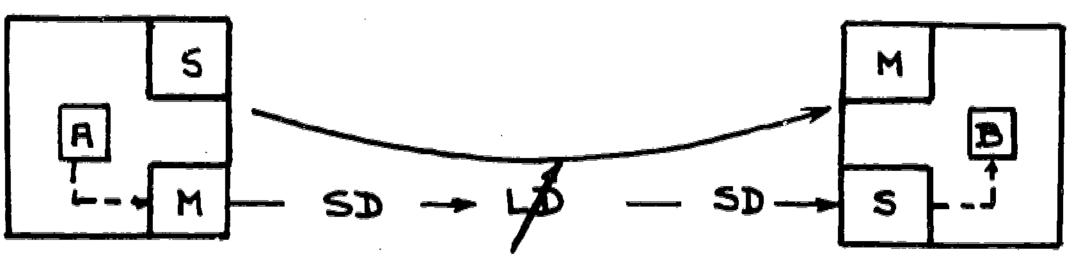
One reason for introducing SD and LD modes is that they enable us to define with some precision the domain of information science, as distinct from that of other sciences. Like all definitions, this one is only as sharp as the concepts (or the classes) used. Consider Fig. 4 as an overall model of three coupled segments, such that the coupling between segments is always in the same relative spatial and temporal order shown. Spatially, it does not matter where recipient B is located, but B is always separated from A by an "isolating" segment, bc . Temporally, A sends the message prior to the time at which B receives it. Two general classes of human communication now arise, depending on whether the message is propagated solely in SD mode, or is transformed into LD mode at some stage in segment bc, and then retransformed to SD mode for terminal coupling with B . This is shown schematically in Fig. 5. With SD-only propagation the main time delays tend to occur at the terminals ("source and sink" in cybernetics; "sender and recipient" in information science). This involves a relatively tight bond. The recipient usually must identify the sender, or at least have a channel that connects back to him. The path becomes a closed loop typical of the conversational



A. BASIC 1-WAY MESSAGE PROPAGATION MODES:
SD
SD— LD—SD



B. SD: USUALLY PERMITS FEEDBACK.
CLOSED LOOP (CONTROL) COMMUNICATION



C SD-LD-SD : USUALLY BLOCKS FEEDBACK.
OPEN LOOP COMMUNICATION

FIG.5 EFFECT OF MODE OF PROPAGATION
ON COUPLING AND CONTROL BY SOURCE.
: VARIABLE TIME DELAY

mode, and also of feedback regulation and control. The closed loop predominates in all society, primitive and modern. Radio, telephone, television and other rapid signal transmission systems tend to retain the small time delays in the central segment, almost essential for closed loop communication. This is the main domain of cybernetics, for without control of the sink by the source possibility of regulation is greatly reduced. On the other hand open loop communication is the principal domain of information science. We now define it in terms of the concepts introduced so far. The proposed domain of information science is simply the set of all three-segment message paths (i.e., initial and final segments are living humans whose signal sending and receiving organs are appropriately matched to the channel of the signal in the central segment, the environment; and sender-environment coupling occurs prior to environment-recipient coupling) such that the mode of signal propagation is usually but not exclusively SD-LD-SD (or briefly, "SLS").

It will be noticed that, since SLS predominates on the "IS path", time intervals between the two couplings of the three segments may be very long. In fact, transduction at b may precede that at c by years, centuries, millenia. The LD message acts as a variable storage device and variable time delay which combines enormous versatility. A third kind of versatility is the "serial addition" of recipients, since the message may be "non-destructively" read out of its LD carrier at various times; and the "multiplication" of recipients through message replication. Thus the LD message enhances man's capacity to carry a message in his mind, lengthens its retention often beyond his lifetime, and extends the number of recipients beyond his capacity to contact them in space and time. If we regard the general goal of a message as "reaching a set of recipients", then the LD mode greatly increases the alternatives open to the sender for goal-fulfillment.

The above generalization of the model of an IS path to that of an IS domain as the set of all such three- segment paths has been criticized from the operational point of view. The concept of the IS domain was shown to be sharpest and clearest for small spatial and temporal intervals between sender A and recipient B. This resembles the increasing difficulty of an operational definition of "simultaneity" in physics.[32] However, by changing the set of operations performed to define the concept, from those of a single observer to those of a social group (scientists), the operational definition can be extended to longer time intervals.[6]

Before discussing the impact of cybernetics on these models, it should be clear from the discussion that cybernetics is the science of control of systems, for which communication turns out to be the means of control; while information science is the science of propagation of meaningful human messages. The two fields are intimately related through concern with communication, but with inverted emphases:

	End	Means
Cybernetics	Control	Communication
Information Science	Communication	Control

V. Application of I-III To IV

Cybernetics makes its main impact on information science through the concepts of variety and its transformations along the IS path, and of requisite variety at all path stages requiring regulation. Now ideas which apply generally to many if not all situations must themselves be simple, general, flexible, without too much structural detail. The concept variety has been shown to enter into all cognitive processes in one way or another. However, the precise way always has to be defined for each particular case.

The law of requisite variety has a simple structure displayed graphically in a diagram of immediate (or of ultimate) effects. Even so, the set {E,D,T,R,C} are not always readily identifiable in applying the diagram to a given situation. In fact, each variable is often multiple in structure, and expressible as a vector.

In applying the law of requisite variety to the three segments of the IS path it soon becomes obvious that the <u>weights</u> of the segments on information scientific phenomena are unequal. By far the most significant phenomena occur in terminal segments ab and cd (I and III).

> The most critical processes in access occur in minds, not in data files, libraries or computers. Access to knowledge requires completing an IS path -- connecting two minds across a variable physical segment.[33]

For this reason we divide this section into two. First we discuss the most difficult subject -- cybernetic applications to the terminal segments of the IS path. Here we are on much less firm (or observable) ground than in discussing the second segment, bc. The terminal segments are among the most difficult and complex regions of investigation of nature -- the regions of human cognitive/affective/motor systems. These systems are of course involved in other sciences.

1. Requisite Variety and IS Path Terminals

In Section I it was mentioned that the concept 'variety' applied to all mental levels of abstraction. These and a theory of communication have been modelled.[3,4,5]

> The mental levels are not hereditary but built up as new structures, by the body. They are constructed from traces of direct sense impressions (DSI's). Stored DSI's were called SSI's, and sets of naturally associated SSI's were named I-sets. Abstractions from I-sets were called A-sets. That is, an abstraction was presented as a single mental structure based on association of a finite number of SSI's. Two main kinds of association were recognized: natural association and conditioned or Pavlovian association. The former gives rise to our concepts of real objects, and through external and internal (body) constraints preserves the original order in space and time of the sampled stimuli. The latter are associations of arbitrary partial patterns -- simpler than the patterns in natural DSI's, yet, because sensed, yielding DSI's. They are symbols, and their role is to be arbitrarily associated with ("attached to") the naturally associated A-sets or concepts.
>
> Physically associated with the concept, the symbol DSI when received acts as a trigger, handle, or switch evoking the abstract concept to which it is conditionally associated out of latent memory into active awareness. Thus in "conversation" a speaker "plays upon the keyboard" of the recipient's A-sets by means of the attached symbols, i.e., (SSI's triggered by) DSI's. For only a stimulus leading to a DSI can cross the inter-individual separation (segment bc).[5]

Sensing and storing sensed data in memory (I-sets) vastly increases the variety of direct experience available to man the self-regulator. He is able to encompass a wider panorama of his environment, a longer time-span than if he were memory-less. Thus we may regard memory storage as a direct application of the law of requisite variety. It gives man (R) more options in the form of greater awareness of disturbances, (D), real or potential, in his environment.

On the other hand, if this were the only cybernetic device at work it would soon come to a halt in the face of an even greater disturbance than ignorance of environment. This would be the inability to respond and act because of the complexity of variety stored. Therefore we look for, and soon find, another device to relieve the flood of sensory data. Using stored sensory data as base, the body forms (non-destructively) new "higher" mental structures. It does so by "abstracting" a greatly reduced number of simplified homomorphisms. That is, the second application of the law of requisite variety is not to increase variety available to R, but to decrease variety in D.

> "... the object of a homomorphic mapping is to make a many-one reduction of variety."[34]

The effect of reducing the complexity of variety to a manageable size, so as to fit within R's limited span of attention and meet other constraints, is to make the V(R)/V(D) ratio larger, and again increase R's versatility. Thus a greatly oversimplified model of cognition consists in two very different transformations of variety, both motivated by the law of requisite variety toward increased versatility:

1) widening the sensory base by expanded storage (+V(R)).
2) contracting and simplifying the sensory base by interposing a set of new structures of homomorphic mappings on the sensory base (-V(D)).

The new structures contain less (or no) sensory variety and more relational variety. This is intuitively what we expect of an abstraction -- we cannot recreate it as a scene, sound, odor...it has little or no "pictorial" content of its own (although, since it is associated with SSI's, it often recalls specific SSI's). Its vague sensory content has less to do with structure than with function. It stores relations defined on the sets of stored SSI's (I-sets).

As described elsewhere[3,4] an abstract concept or abstraction is also a class. A class is a mental construct which has two aspects,

1) a qualitative "decision" rule for class membership or non-membership (intension
2) a quantitative number of members, (the class extension). In the model below, the members are SSI's, and all higher classes are formed from them:

We have conjectured the existence of structures in the brain (call them A-sets) derived from naturally associated sets of stored sense impressions (I-sets). Each member of an A-set has the nature of a "property" of the I-sets. It is also the intension which defines a class. For the special case of an A-set based on tree-like I-sets we suggested a simple model, giving the minimum number of naturally-associated SSI's required to form the i^{th} "level":

$$N(i) = 2^{i-1} \qquad (i=1,2,\ldots,m)$$

where m is the highest level constructed within the brain on that I-set base. Then the total number of naturally associated SSI's would be at least

$$N(m) = \sum_{i=1}^{m} n(i) \qquad n(i) = 1+2+4+8+\cdots+2^{m-1}$$

$$= 2^m - 1$$

In this model each successive level contains one more naturally associated SSI than the entire sum of all those in lower levels. What interpretation can be placed on this rapid increase in minimum class extension? Again there is a straightforward cybernetic interpretation: the levels represent decreasing variety of class intension (or partial-patterns) in the set of SSI's which define the class, i.e., decreasing contributions to the homomorphisms, from each constituent SSI. A decrease in variety may be regarded in several ways. Most obviously, it represents a lessening of constraints, so that more objects can be found that comply with it as a defining rule for a class, i.e., minimum class extension increases. Another aspect is that, if range in variety used to define a field is narrowed, the range in the corresponding concepts is narrowed, or the concept stability increases. That is, concept stability would be some inverse function of permissible range in variety. But class extension is such an inverse function (although not necessarily the correct function). Hence we may regard the increasing extension and decreasing intension in the abstractions as a measure (of some sort) of increasing conceptual stability. Finally, there is another aspect of the same phenomenon. The decrease in class intension also corresponds to increase in versatility of response behavior on the part of the person whose brain is involved. There are innumerable examples. For instance, when John is five, the question "how many are two cows and three horses?" is no poser, for he does not see the difficulty. As he matures, he senses that cows and horses differ and cannot be added. When he has developed still more levels, he calmly uses an upper level, and answers that "two domestic animals and three domestic animals are five domestic animals." Thus he has acquired additivity by use of a higher, more abstract level. The physicist does the same when he insists on "dimensional homogeniety" in terms of his equations. The mathematician strips all the internal structure of intension from his mental objects, leaving only their externally discriminable structure -- their number and order. He thus creates the integers, and thereby gains still more versatility. Greater generality is equivalent to less defining intension of a class, to less variety. And the rule is that, the more general the class, the more "particular" cases it subsumes, and the more versatile is the mental owner in performing mental operations.

Now let us look in the opposite direction. At the bottom level is the SSI. It contains the maximum variety we can experience in one observation. It has the least versatility. Using only the SSI we cannot communicate internally in thought, or externally, in a message. If we look a few levels higher, we still cannot communicate. For example, we cannot communicate level 2. The uniqueness is still too great, i.e., there exist no comparable structures in the mind of another person. However, at a certain level of stability the conceptual structures in two minds begin to be sufficiently similar so that the indication of the concept structure by one person finds something similar which can respond, in the other person. When this hypothesized threshold of stability for communication is reached, there is an enormous increase in versatility. For now the two persons can function as a social unit, mutually assisting eachother's goals. This hypothetical level we indicate by n(c), the minimum number of naturally-associated SSI's for the threshold conceptual stability necessary for interpersonal communication. Evidence for the existence of such a threshold are the facts that we cannot communicate DSI's and SSI's, and that we can and do communicate by signs or symbols that evoke abstract classes. For levels higher than c, communication becomes easier and easier. The probability that our conditioned associations (symbols) evoke abstract concepts (if they are present) becomes better and better. This idea underlies the ease of communication by small groups of professionals who share the same sets of abstractions. The precision of mathematical concepts, for example, is no accident. The fact that they can be (not necessarily always are) so precise is attributable to their enormous suppression of variety. They are actually derived from very large I-sets of which the mathematician loses awareness. In fact the difference between mathematics and other sciences lies in a kind of superversatility -- the mathematical concept structure is not necessarily reconstructed so as to embody the constraints on patterns of variety observed in the real world of DSI's. Yet one of these constraints persists in a way that the mathematician must observe. He uses it as his link with "empirical reality". The constraint on his synthetic patterns is that of validity of proof. Proofs allow enormous simplication (a homomorphic device) since they permit suppressing the variety in lower levels of abstraction, and retaining only the reduced patterns at the higher levels. A proof is a rule for interlevel transitions, "inference", in logic. The example below applies only to levels of abstraction based on tree-like I-sets, as in the simplified model:

Hypothesis 1:	All A is B	(if A then B)
Hypothesis 2:	All B is C	(if B then C)
Conclusion :	All A is C	(if A then C)

Assuming the mathematician can demonstrate the truth of the conjunction of the hypotheses, then a valid conclusion follows, and the middle term B is unnecessary, as in fact is A, since the whole process is designed to show that A also conforms to rule C. This type of constraint, empirical in origin, has the versatility of applying to any three levels. The example merely suggests that logic too is a device for reducing variety, for regulating and controlling it, i.e., for increasing the versatility of its user.[4]

In the discussion so far it has been suggested that the law of requisite variety applies to the gathering of large quantized units of sensory variety (DSI's) stored (as SSI's) in memory banks (I-sets) which preserve the combined constraints of body and environment. In order to use these data other processes are needed. One of these is the capability, given an incident DSI, to extract features, recognize them and classify them at once so as to endow the DSI with "meaning". It has been suggested that this deductive pattern recognition is accomplished by the A-sets once they are built up in the body inductively. A-sets are partial patterns. They act like prisms or diffraction gratings on a beam of white light, instantaneously decomposing and analyzing an afferent DSI into its "A-set spectrum" of "characteristics".[35] No mechanism is suggested. However, we know this process occurs. We do rapidly understand a DSI "in light of of experience". Understanding is not sensory, but relational. The isolated DSI or its memory trace SSI has no meaning <u>by itself</u>. It <u>acquires</u> meaning in relation to mental structures previously formed. This power to "deduce", on which our expectations and predictions are based, represents an enormous gain in versatility. It enables man the self-regulator to conduct self-advantageous action.

Two other homomorphic transformations on variety are needed for a decision. The first might be called cognitive selection, a binary mapping of those alternatives which seem most relevant (in terms of a scan of the sets of spectra of characteristics) into the narrow channel of immediate awareness. The second transformation might be called selective affective response. It is an ordering of the cognitively selected alternatives by feelings and drives - decision by self-utility or value. Finally, the motor discharge in segment I (ab) innovates responsive action.

All of these processes are immensely complex. However, looked at as a whole they simply increase the capability of the body for self-regulation. They can be interpreted as applications of the cybernetic law of requisite variety.

This discussion renders somewhat more precise our opening remarks. The most significant information scientific phenomena do take place in the terminal segments of the IS path, but the third segment (cd) is the more complex of the two. It performs the main functions of sensory observation, data storage, data analysis and compression, cognitive selection. Affective decision is not necessarily localized. It may take place "between" segments III and I. Segment I has comparatively few functions -- those for propagating the coded results of decisions out of the body as signals (stimuli, symbols), across segment II, the "insulating gap" between minds.

We also recognize the logical and practical impossibility of discussing what happens in the central segment without also discussing what happens in the terminal segments. In nearly every case, all three segments are engaged. The temporal and spatial orders are preserved and remain invariant. Therefore we cannot segregate these operations as long as we consider them as parts of one meaningful whole. A possible exception might be use of segments II and III in "passive observation" -- observing nature. It cannot occur in observing artifacts (such as messages). For even if, in passive reading, we are using segment II (a book) and III (visual-mental tract) we are <u>also</u> completing as IS path -- initiated in some segment I. This "holistic" view has been used for a simple graph-theoretical demonstration that an IS path has maximal length (therefore can be treated as uniquely defined), <u>not</u> in terms of a physical space-time interval, but in <u>numbers of stages, or transformations of variety</u>. Unique and exhaustive, the IS path can serve as a convenient reference frame for two-person communication. It allows us to describe the information scientific phenomena of pairs (sender and recipient). With due caution in making a new operational definition the concept can be generalized to its societal (scientific) generalization, the collection of IS paths or IS domain.[36]

2. Requisite Variety and IS Path - Segment II

In view of the practical inseparability of the three segments we examine from the viewpoint of cybernetics models of information scientific operations which depend typically on artifacts in segment II. The first is that of search of a collection. The searcher embodies segments I and III, the collection is in segment II.[37,38]

> The librarian and information processor perform many professional tasks by reducing the variety among documents and other artifacts used in common by society. For example, they construct categories of inclusion or hierarchies (precoordinated); or they may set up the potentiality of such hierarchies, or of simple class conjunctions (post-coordinated). The use of precoordination ("classification") in information search was studied in the model:

$$N = C\ K\ V\ T$$

N represents the number of units (e.g., words) in the collection, T the time spent sensing (seconds), V the average rate of sensing (words/second), K the ratio of homomorphic compression, and C the ratio of units of the original collection to the units still to be sensed. That is, V represents the process of sensing and widening the sensory base of stored variety (reading). K represents the compression ratio when, e.g., a book with 100,000 words is abstracted to form a new "surrogate" of itself such as an abstract or a catalog card containing, say, 100 words. Then $K = \frac{100,000}{100} = 1000$. This is the numerical but not semantic measure of homomorphic compression.

If the collection contains 1,000,000 volumes and the selection process reduces this to 100,000 volumes, the value of C would be $\frac{1,000,000}{100,000} = 10$.

K and C respectively reflect two of the three suggested basic mental processes: K, homomorphic simplification and suppression of variety; C, isomorphic sensing of variety, coupled with its elimination by selection. K and C can both be regarded as homomorphic mapping processes, one by compression into a simpler structure with retention of certain invariants (the meaning), C by mapping an original structure into a binary function (1,0). The result of the latter is to select parts of the original without otherwise altering them.[38]

Since, as mentioned, V represents a widening of the sensory base of variety, the model simulates three main cognitive processes used in search of collections. In information search there is a reciprocal use of K and C such that the total number of symbols sensed (VT) is kept small, and search takes place on the compressed and selected surrogate collection (N/CK). Figure 6 shows an "inverse homomorphic" use of a classification system.

During search C <u>increases</u> from 1 to N (the end of the search, when the last word is about to be sensed); while K has one or more constant values, which <u>decrease</u> until K = 1. In all cases the manipulation of K and C follows a certain order. Since the act of selecting requires variety of more than 1, the values of K must be reduced <u>before</u> the value of C can be increased. For example, suppose a system is used (somewhat like the Deway Decimal System) in which the collection is first divided into ten parts. If the entire collection were first represented by one word, say the word "Collection", then the value of K would be, initially

$$K_0 = \frac{N}{1} = N$$

and no selection could be made from the one word. If however, K were reduced so that (decreasing K by a factor of ten)

$$K_1 = \frac{N}{10}$$

then it would be possible to select one or more of the ten words. Thus the variety in the compressed collection had to be <u>increased</u> from 1 to 10, <u>before</u> the selection from a variety of 10 could be made. In any decision process, there must be a variety of at least 2. In information search with a classification system we alternate decompression (the inverse of the homomorphic compression that took place when the more abstract class was formed), with selection. We go from the abstract to the concrete, always keeping the amount of variety to be handled at each stage small (and therefore the sensed message M short). In this way the total sensed message:

$$\begin{aligned} M = VT &= VT_1 + VT_2 + \ldots + VT_m \\ &= M_1 + M_2 + \ldots + M_m \end{aligned}$$

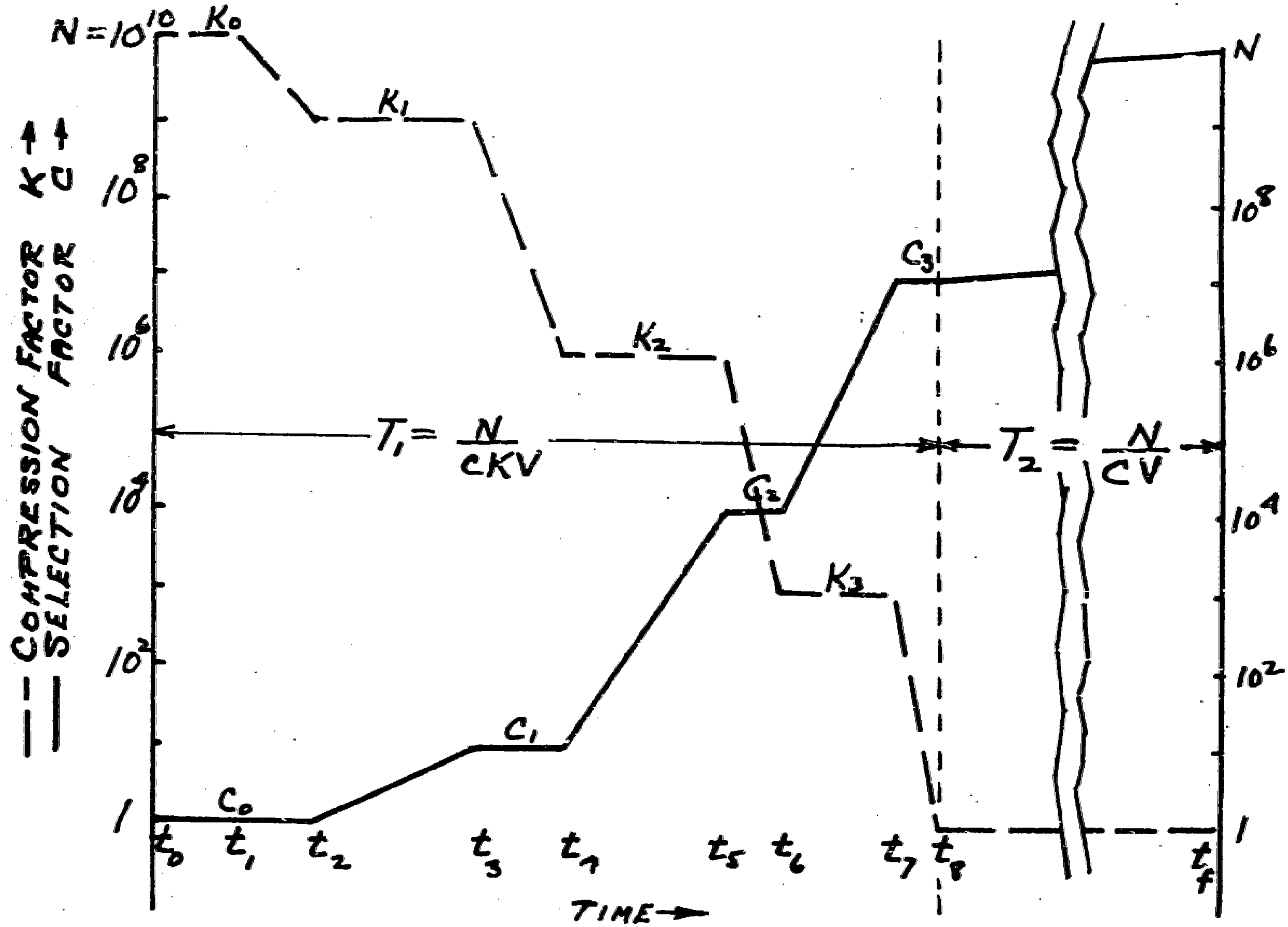


FIG 6 ALTERNATIONS OF K AND C BLOCK VARIETY IN HIERARCHIC SEARCH

where m corresponds to the lowest value of K greater than 1. At K = 1 the nature of the search changes from decision-making (selecting what to sense) performed on the homomorphically-mapped image-collection, to sensing the selected part N/C of the original collection. Advantage can be taken of this change in function, as described below. The successive reductions in K permit the selection process to take place rapidly (since at each stage j only the few words M_j are sensed) and still preserve exhaustive coverage. The homomorphic compression represents the whole remaining collection at each stage. Note that in the schematic search diagram K and C change inversely and alternately with K preceding C. The alternation ceases after K = 1 and thereafter the rate of climb of C is much slower, since the "reduction" of the collection now occurs only at the ordinary reading rate V. The entire time of search can be symbolized by

$$VT = VT(K > 1) + VT(K = 1) = M(K > 1) + M(K = 1) = M$$

in which the length of M and T are controlled by regulating the amount of variety used.[38]

Another example is that of an indexer using a thesaurus (figure 7).[39] Here the disturbance D is the variety in the set of uncontrolled vocabulary terms in the literature to be indexed. This variety "threatens" the correctly indexed or "controlled" vocabulary terms E. The variety in D is suppressed to the lower variety in the controlled vocabulary by indexer T, whose selection is jointly dominated by D and thesaurus R.

A third example is provided by a reference librarian who regulates the interaction of a library user and the library. The user "threatens" (his or her own) access to the library's information by his or her variety in request for information or in approach to the system of catalogs, stacks and collection. The reference librarian regulates this variety by making proper selections following the user's query. The reference librarian in turn is "controlled" by the philosophy or policy of the library (Figure 8).[40]

Another example is that of a classification system (Figure 9a)[41]; such as the Dewey Decimal Classification System (DDC) or the Universal Decimal Classification System (UDC) or Colon Classification System (Figure 9b). The variety in a new concept to be classified (D) "threatens" the set of

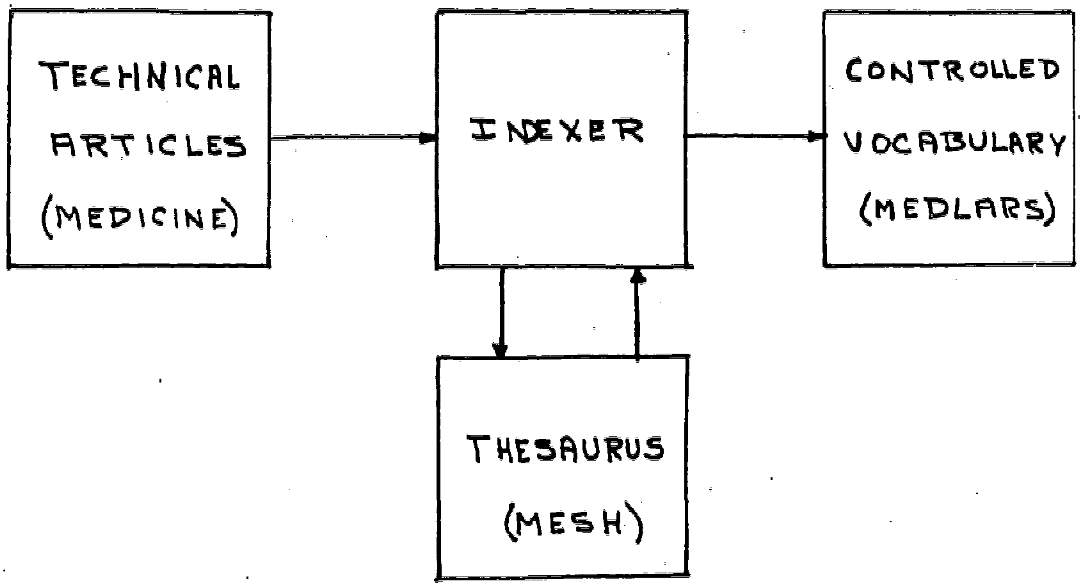


FIG. 7. THESAURUS AS REGULATOR: DTR

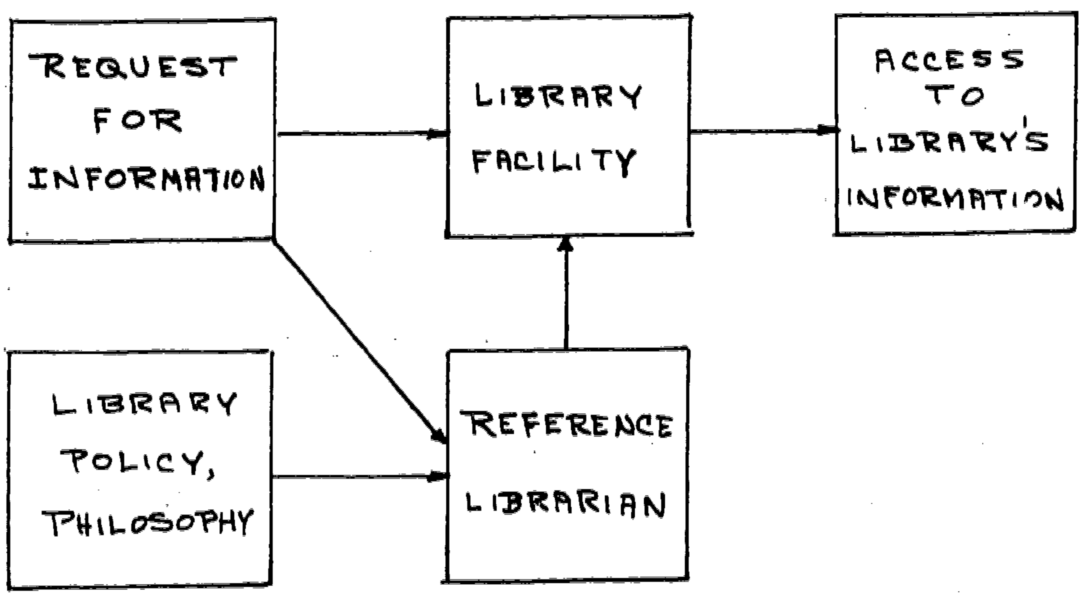


FIG. 8 REFERENCE LIBRARIAN AS REGULATOR: DR, DTR

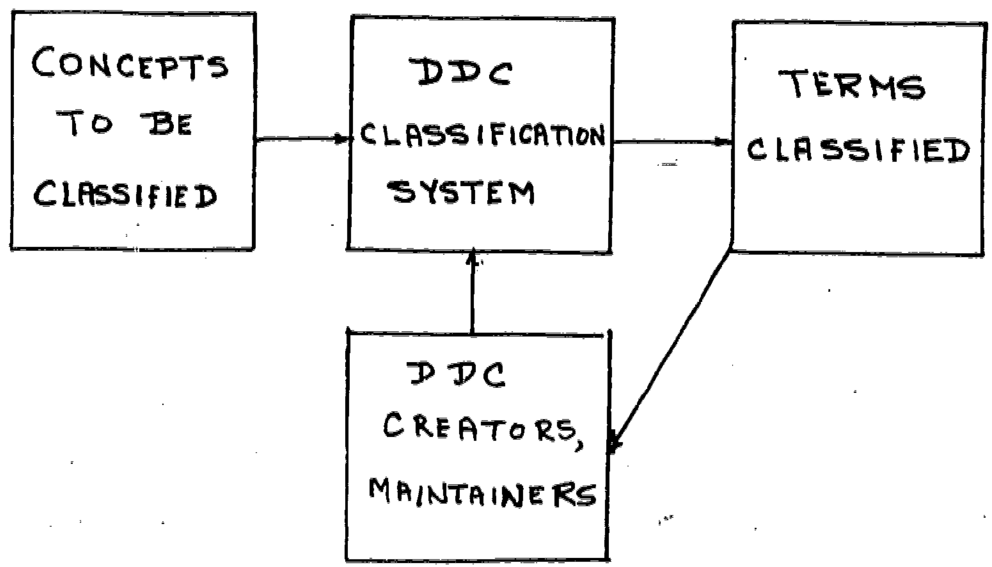


FIG. 9a. MAINTENANCE OF DDC CLASSIFICATION SCHEME AS REGULATOR: D T E R

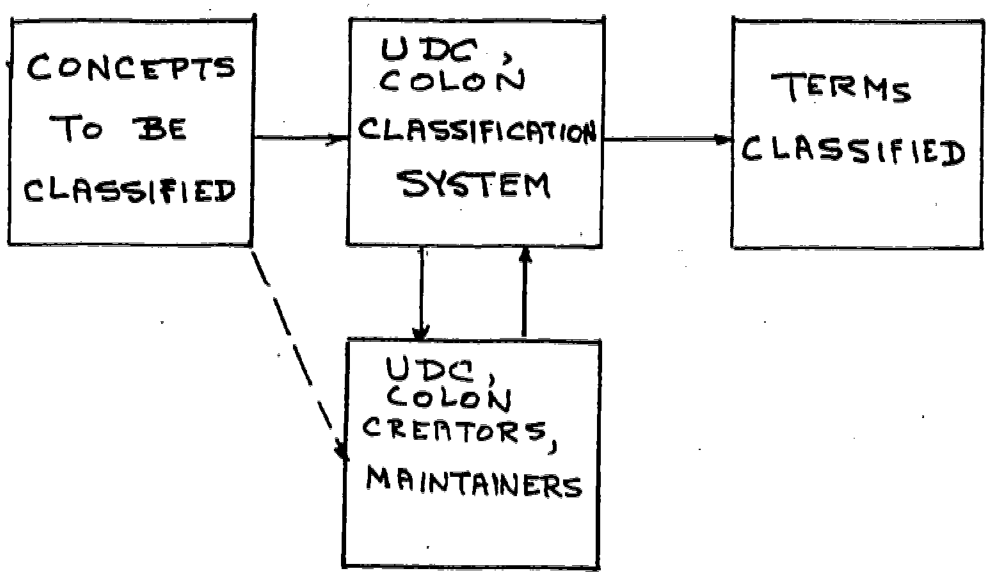


FIG. 9b. MAINTENANCE OF UDC, COLON CLASSIFICATION SCHEME AS REGULATOR: D R, D T R

correctly classified concepts (E). D is regulated by the creators and maintainers (R) of classification scheme T.

The various organizations for maintaining knowledge classification schemes are not equally effective as regulators. In DDC (Figure 9a), the regulation is by error correction, i.e., the message path from D to R is DTER. In UDC and Colon Classification (Figure 9b) the organization of maintainers is such that the path is either DTR or DR.

A final example is that of an academic acquisitions system (Figure 10).[42]

The system E to be protected is the set of orders (E_+ successfully, E_- unsuccessfully) filled. The disturbance D is the set of academic orders, the environment (or "table") T is the acquisitions department and vendors, the regulator R the acquisitions librarian, and C the fund control against overexpenditures. The path is DTR or DTER.

In conclusion, this paper has briefly outlined the cybernetic concepts of variety, its transformation, and use in regulation through the law of requisite variety. To apply these ideas to information science two "framework" models were sketched, the IS path and IS domain. Finally, examples of phenomena selected in all three IS path segments suggested how the law of requisite variety applies almost universally to transformations on the IS path and within the IS domain. Each example differed structurally but was functionally isomorphic. The isomorphism is that of a goal - equilibrium - disturbance configuration. Moreover, in IS the end is always communication and the means regulation of threatening variety. This systematically recurring pattern of a goal-seeking type provides an initial theoretical base for information science. "Initial" because the law of requisite variety is abstract and its constraint functional. We cannot expect it to predict detailed structural constraints. It applies to too many situations--not only to regulation in information science but to all regulation, with different goals and different

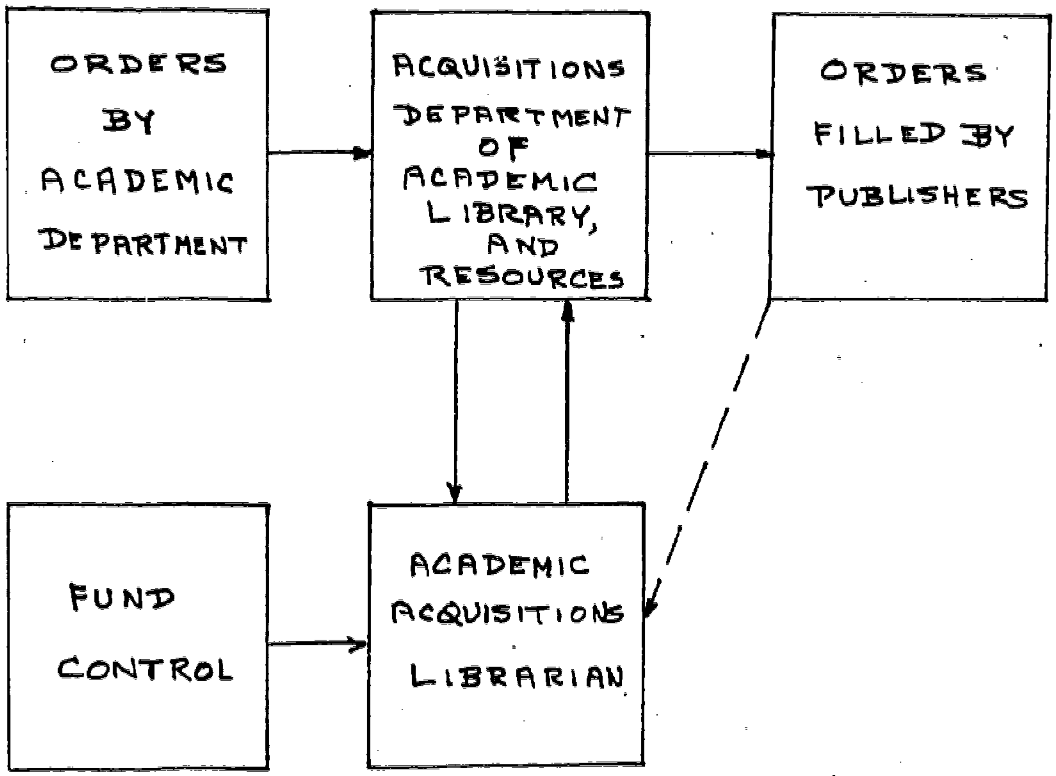


FIG 10. LIBRARIAN AS REGULATOR of ACADEMIC ACQUISITIONS: DTR, D T E R

equilibrium states. On the other hand we may expect such a broad law to lead to further discovery. It should help us discover and formalize an eventually multidisciplinary theory of information science by pointing not only to the wealth of cybernetic connections but to the fundamental contributions of other sciences.[43]

References and Notes

1. Wiener, N. Cybernetics or Control and Communication in the Animal and the Machine. The Technology Press. Wiley, 1948.

2. Ashby, W.R. An Introduction To Cybernetics. University Paperbacks, 1956, reprinted 1968.

3. Heilprin, L.B. "Information Storage and Retrieval as a Switching System", Switching Theory in Space Technology, Aiken and Main, eds., Stanford University Press, 1963, 298-332.

4. Heilprin, L.B. "On Access to Knowledge in the Social Sciences and Humanities, from the Viewpoint of Cybernetics and Information Science". To be published by the Library Science Department, Queens College of the City of New York, in proceedings of a Conference on Access to Knowledge and Information in the Social Sciences and Humanties: Problems and Implications, April 5-6, 1972, New York City.

5. Heilprin, L.B. Impact of Cybernetics on Information Science, and Vice Versa. Paper for FID Budapest Conference, FID/TM Meeting, 5 September 1972. In Systems, Cybernetics and Information Networks, K. Samuelson, Ed., FID/TM, Stockholm, Sweden, 1972, pp. 22-33.

6. Heilprin, L.B. Guidelines Toward Operational Definitions in Information Science. Paper for NATO Advanced Study Institute in Information Science. To be published 1973.

7. Shannon, C.E. and Weaver, B. The Mathematical Theory of Communication. U. of Illinois Press, 1949, p. 5.

8. Ashby, W.R. Loc. cit., Chapter 7, 120-127.

9. Woodward, P.M. Probability and Information Theory, with Applications To Radar. McGraw-Hill, 1953, Chapter 3, pp. 43-49.

10. Ashby, W.R. Loc. cit., Chapter 7, pp. 127-132.

11. Ashby, W.R. Loc. cit., Chapter 7, pp. 132-134.

12. Heilprin, L.B. Reference 4.

13. Beer, S. Decision and Control. Wiley, 1966, pp. 239-240.

14. Ashby, W.R. Loc. cit., Chapters 11, 12.

15. Abramson, N. Information Theory and Coding. McGraw-Sill, 1963, Chapter 2.

16. Woodward, T.M. Loc. cit. Ref. 9.

17. Hartley, R.V.L. "Transmission of Information". Bell System Technical Journal, July 1928.

18. Ashby, W.R. Loc. cit., Chapter 9, pp. 174-175.

19. Heilprin, L.B. Reference 5, Section 7.

20. Equivalent Statements: A implies B, A is sufficient for B, B is necessary for A, A only if B, A is included in B, A is part of its environment B.

21. Heilprin, L.B. Reference 4, Section I.

22. Ashby, W.R. Loc. cit., Chapter 2,3.

23. Ashby. W.R. Loc. cit., Chapter 4,5.

24. Ashby, W.R. Loc. cit., Chapters 4,8,13.

25. Simon, H.A. The Sciences of the Artificial. MIT Press, 1969.

26. Ashby, W.R. Loc. cit., Chapter 4.

27. Ashby, W.R. Loc. cit., Chapter 11, p. 207.

28. Ashby, W.R. Loc. cit., Chapter 11, p. 211.

29. Ashby, W.R. Loc. cit., Chapter 11, pp. 207-11.

30. Heilprin, L.B. "Toward a Definition of Information Science", Automation and Scientific Communication, Proceedings 26th Annual Meeting, American Documentation Institute, Chicago, Ill., October 1963, 239-241.

31. Heilprin, L.B. "On the Information Problem Ahead", American Documentation, 12, January, 1961, 6-14.

32. Bridgman, P.W. The Logic of Modern Physics. MacMillan, 1932.

33. Heilprin, L.B. Reference 4, Abstract.

34. Beer, S. Loc. cit., p. 257, Reference 13.

35. Heilprin, L.B. Reference 4, Section II.

36. Heilprin, L.B. Reference 6, Section 3.

37. Heilprin, L.B. and Goodman, F.L. "Analogy Between Information Retrieval and Education", American Documentation 16: July 1965, 163-169.

38. Heilprin, L.B. Reference 4, Section 6, and Reference 5, Section 8.

39. Al Naib, M. MEDLARS As An Information System. Term Paper, LBSC 721, University of Maryland, 1972.

40. Roderer, N.K. The Library as a Regulator: Two Models for Discussion. Term Paper, CMSC 737, University of Maryland, 1972.

41. Goldstein, D.M. A Cybernetic View of Classification, Term Paper, LBSC 721, University of Maryland, 1972.

42. Meszaroz, I. A Systems Approach to Out-of-Print Acquisitions for Academic Libraries, Term Paper, LBSC 721, University of Maryland, 1972.

43. Thanks are due Ms. Shelly Rowe, D. Smith, and B. Wahl for unsparing use of their time and superb typing.